2022 9th International Conference on Future Internet of Things and Cloud (FiCloud) | 978-1-6654-9350-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/FiCloud57274.2022.00041

A Data Mining System for Enhancing Profit Growth based on RFM and CLV

Lahcen ABIDAR, Ikram EL ASRI, Dounia ZAIDOUNI, Abdeslam ENNOUAARY
INPT, Rabat, Morocco
abidar.lahcen@inpt.ac.ma*, [elasri,zaidouni,abdeslam]@inpt.ac.ma

***Abstract*—Understanding customer has piqued the curiosity of industry experts, marketing executive's and academics researchers. The potential value of a customer to a business can be a key factor in decision making. Customers assessment, identifying the differences between them, and ranking them is one of the major issues in customer-based enterprises. However, leveraging all of our strengths, as well as new technologies such as machine learning algorithms and data treatment, we can now build very powerful models that allows us to better understand consumer requirements and behaviors, and respond properly to meet his demands. In the present paper, we propose a model based on Customer Lifetime Value (CLV) and the Recency, Frequency and Monetary (RFM) model to address these concerns. This model will assist us in gaining a clear understanding of our consumer categories and determining what actions we can take for each category.**

***Index Terms*—Data Mining, Customers segmentation, CLV, RFM model, Machine Learning.**

I. Introduction

Clustering before analysis allows a company to see patterns clearly across the life-cycle of a customer (or user), rather than slicing across all customers blindly without accounting for the natural cycle that a customer undergoes. By seeing these patterns of time, a company can adapt and tailor its service to those specific studies. CLVs help us determine how much we should spend to acquire customers by estimating how much value they will bring to our business over time. We will be able to understand which customers to focus on, and more importantly, why we should focus on them, not just running around to make ends meet. Customer lifetime value is a clear view of the benefits of acquiring and retaining any particular customer. Not all clients are the same. Knowing our CLV has many main benefits, It encourages repeat sales and revenue. CLV finds customers who spend more in store. It can help to understand which products they love and which products make their lives better. we can use CLV to track the number of sales per customer and develop strategies to increase repeat purchases and profit margins. For every 5% increase in customer retention, sales can increase by up to 95% [1]. Knowing the customers CLV helps us to improve the loyalty. The strategies we use to increase CLV can improve customer support, products, referrals, and loyalty programs, resulting in more repeat customers and higher retention rates. Existing customers buy more frequently and spend more than new customers. It reduces the ratio of Lifetime Value (LTV) to Customer Acquisition Cost (CAC). Some research shows that the average cost of customer acquisition ranges from $127 to $462, depending on the industry [2]. The higher LTV is relative to CAC, the faster company can grow. By increasing customer lifetime value, we can create benchmarks on how marketing impacts customer profitability. However, the required short-term marketing investment in a strategy is also critical because cost pressure on marketing budgets is increasing, a trend that has accelerated during the COVID-19 economic downturn [3]. It's important also to assess the power of each possible targeting strategy by providing a preliminary comparison of their effectiveness in terms of both short-term marketing campaign cost and long-term CLV increase.

II. Related Work

Customer lifetime value (CLV) is a metric that measures the value of a customer over the course of his relationship with the company. This metric is extremely important and is widely used by a wide range of businesses, including financial institutions, retail stores, telecommunications companies, and others. Several models for calculating customer lifetime value have been proposed. Kumar et al. (2008) [4] attempted to answer the question "Which customers should be targeted?". Marketing investments made after segmentation demonstrate a significant future contribution to the firm.

Glady and Baesens (2009) [5] worked in the banking sector. They use Berger and Nasr (1998) Model and data provided by a finance corporation in Belgium to demonstrate the benefit of the new version of Pareto/NBD analysis by adding some stuff to the previous one. It displays data from 2000 to 2005, with a total of 460.566 customers. The Pareto/NBD model has been highly successful as a tool for customer base analysis it aims to model whether or not customers are alive and, if alive, how frequently they purchase. The initial model was proposed by Schmittlein et. al. (1987) [6]. Hwang, Jung (2004) [7] collaborated to create a new CLV method and create a segmentation based on CLV. This study used data from a Korean telecommunication corporation to create a new CLV model that included the customer's historical prospective contribution and segmentation. 16.384 customers have 200 different variables. Only 2000 of them were chosen for this investigation. Gloy and Akridge (1997) [8] made an attempt. To demonstrate the value of CLV in determining marketing strategies in the petroleum industry, they used the Berger

978-1-6654-9350-5/22/$31.00 ©2022 IEEE
DOI 10.1109/FiCloud57274.2022.00041

		Methods		
		K-Means	RFM	Number of Cluster
Authors	**Year**			
Abidar L, Zaidouni D, Ennouaary A	2020	Yes	Yes	Elbow method
Ezenkwu, C.P., Ozuomba, S. and Kalu, C.	2015	Yes	No	Default k
cali, A.,Boyaci, A., and Baynal, K.	2015	Yes	No	The smallest sum of squared error value
Khajvand, M., and Tarokh,M.J.	2011	Yes	Yes	The Largest value of Dunn Index
Yuliari, N.P.P., Putra, I.K.G.D., and Rusjayanti, N.K.D.	2015	No	Yes	MPC (Modif ied Partition Coefficient)
Cheng, C.H., and Chen, Y.S.	2009	Yes	Yes	Default k
Hosseini, S.M.S, Maleki, A., and Gholamian, M.R.	2010	Yes	Yes	The smallest Davies-Bouldin Index value
Hamdi, K., and Zamiri, A.	2016	Yes	Yes	Default k
Lanjewar, R., and Yadav, O.P.	2013	Yes	No	Default k
Kashwan, K.R., and Velu, C. M.	2013	Yes	No	Default k
Dhandayudam, P., and Krishnamurthi,I.	2013	Yes	Yes	Default k

TABLE I: COMPILATION OF CUSTOMER SEGMENTATION RESEARCHERS

and Nasr (1998) model as well as data from a petroleum company. Because different client groups necessitate different marketing mixes, segmentation is critical to marketing strategy. Many researchers have attempted to create the finest customer segmentation they can. Table I shows some compilation of customers segmentation, as well as the methods they employ. The majority of the developments in the literature, focus on modifying and improving the Pareto/NBD and BG/NBD models, which continue to be the performance benchmarks for modeling the consumer repeat buying process. [9], and over the last two decades, there has been a significant increase in interest in targeted marketing research [10] [11] [12] [13] [14]. The present paper proposes a new model Based on RFM and CLV. Using unsupervised machine learning techniques (Clustering). The model analyses customers' transactional data and proposes a set of recommended actions for every segment to enhance the company's profit.

III. Workflow Model

This research will use several approaches and procedures in order to arrive at a final outcome. The resulting framework workflow, which includes customer segmentation, RFM parameters, clustering, data analytics, and targeted actions, is represented in Figure1.

A. Data preprocessing

1) Data Cleaning: There are two general ways to handle missing values in building operational data. The first is to simply discard data samples with missing values as most data mining algorithms cannot handle data with missing values. Such method is only applicable when the proportion of missing values are insignificant. The second is to apply missing value imputation methods to replace missing data with inferenced values. For outlier detection there are two methods: statistical and clustering based methods [15].

2) Data Reduction: Data reduction is typically conducted in two directions, i.e., row- wise for data sample reduction and column-wise for data variable reduction. Various data sampling techniques can be applied for row-wise data reduction, such as random and stratified sampling. Unlike feature selection which only selects useful features from existing variables, feature extraction aims to construct new features based on linear or nonlinear combinations of existing variables [15].

3) Data Scaling: Data scaling is often needed to ensure the validity of predictive modeling, especially when the input variables have different scales. The max-min normalisation (i.e., $x' = x - x_{min}/x_{max} - x_{min}$) and z-score standardization (i.e., $x' = x - \mu/\sigma$) are two of the most widely used methods in the building field, where x_{min} and x_{max} refer to the minimum and maximum of variable x, values of the variable, μ is the mean and σ is the standard deviation [15].

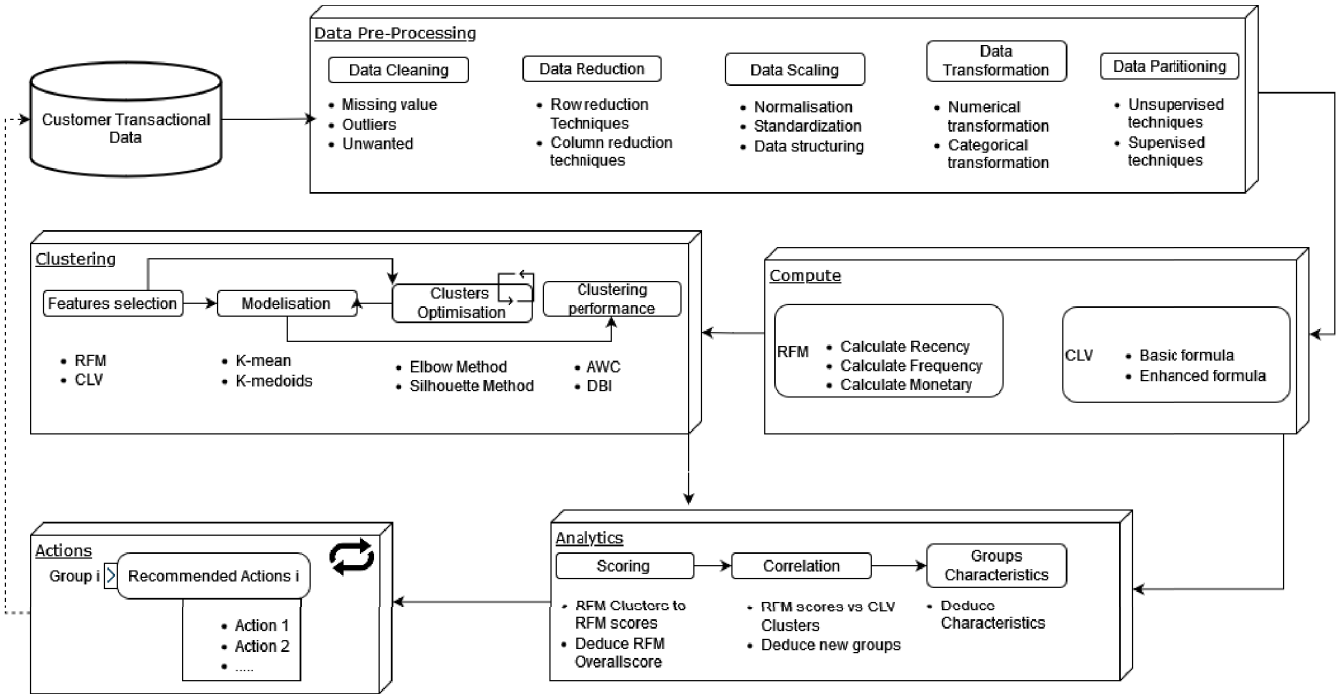
4) Data Transformation: In the building field, data transformation is mainly used to transform numerical data into categorical data to ensure the compatibility with data mining algorithms. The equal-width and equal-frequency methods have been widely used due to their simplicity [15].

5) Data Partitioning: Data partitioning aims to divide the whole data into several groups for in-depth analysis. Clustering analysis and the decision tree methods have been widely used in the building field for this purpose. A number of clustering algorithms have been applied for data partitioning, such as k-means, hierarchical clustering, entropy weighting k-means (EWKM), and fuzzy c-means clustering [15].

B. Clustering

1) Features selection: In practice, it is rare that all of the variables in a dataset are useful for building a machine learning model. Adding redundant variables reduces the model's generalization capability and may also reduce a classifier's overall accuracy. Furthermore, adding more variables to a model increases the model's overall complexity. In machine learning, the goal of feature selection is to find the best set of features that allows one to build useful models of studied phenomena. There are two types of feature selection techniques in machine learning: supervised and unsupervised techniques.

2) Modelisation: A machine learning model is a file that has been trained to recognize different patterns. we train a model on a set of data by providing it with an algorithm that it can use to reason about and learn from that data. Once the



(a)

Fig. 1: Framework Workflow

model has been trained, we can use it to reason over data that it hasn't seen before and make predictions about it.

3) Cluster Optimisation: Every clustering algorithm has its own strengths and weaknesses, due to the complexity of information. The weaknesses of clustering algorithms include determining the number of clusters based on assumptions and relying heavily on the initial selection of centroids to overcome these weaknesses. It is necessary to optimize, one of the popular cluster optimization methods is the Elbow method.

4) Clustering performance: One fundamental question that need to be addressed in any typical clustering system is : how real or good is the clustering. The Silhouette Score and Silhouette Plot are used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters. This measure has a range of [-1, 1] and is a great tool to visually inspect the similarities within clusters and differences across clusters.

C. Compute

1) RFM Compute: RFM is a method used to give every customer a significant value. It is mostly used in marketing and has received particular attention in retail and industry services. RFM is based on the following dimensions 2:

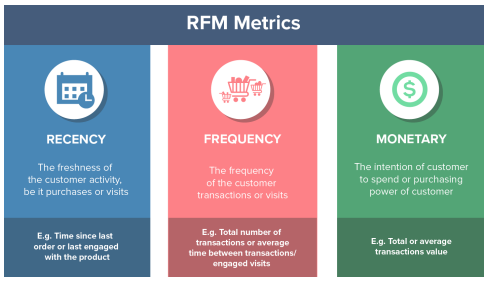


Fig. 2: Recency, Frequency and Monetary Metrics.

2) CLV Compute: CLV is calculated as the sum of the net cash flows from the customers considering the time value of money throughout the expected life of customers. This model can be expressed as in the following formula [16] and table IIis showing it's parameters:

$$CLV = \sum_{i=1}^{n} \frac{R_i - C_i}{(1+d)^{i-0.5}} \quad (1)$$

IV. Case Study

A. Data

The data used in this research was collected from an online retail company [17]. The dataset includes the customer variables which belong to the period from the end of 2009 to

Var	Explanation	Operationalisation
n	Expected life of a customer	n= the total number of periods of projected life of the customer under consideration
Ci	The total cost of customer in period i	Total cost of generating the revenue R i in period i
Ri	Total revenue of customer in period i	The revenues of customers were assigned as their monetary values.
d	Discount rate (annual)	Discount

TABLE II: CLV formula parameters

November 2011. The dataset consist of 16759 invoices about 3881 products made by 889 customers.

An exploratory data analysis revealed the following insights :

- The average quantity of products purchased by a customer is 1079.0
- The average revenue generated per customer is 1797.46
- The average product quantity sold per transaction is 12.0
- The average revenue generated per transaction: 20.17

Now that we have developed a basic idea about how our retail data looks like, let us proceed in the first step of our model data pre-processing .

B. Data pre-processing

1) Data cleaning: Data cleaning process have been executed and some missing values, wrong values have been excluded from data set. Table III shows the attributes used in this study. Table IV shows some corrections we apply to data to make it more clean.

Attributes	Description
InvoiceNo	Unique ID to identify each Invoice
StockCode	Unique ID for each item in stock
Description	A short description for each item
Quantity	Number of items bought
UnitPrice	The price of each item
CustomerID	Unique ID for each Customer
Country	The country were the Customer lives

TABLE III: Attributes

Problem	Solution
Invoice with null values	Remove from the dataset (Not useful for this study)
UnitPrice with negative value	Remove from the dataset (added by the company to Adjust bad debt)
invoice with customerID is not available	Remove rows where customerID are NA since we are going to do customer segmentation.

TABLE IV: Data Cleaning

We use whole prepared population in the analysis. Thus, we did not use any sampling method.

2) Data Selection: for this study we selected the following attributes : Quantity, InvoiceDate, UnitPrice, CustomerID. Those attributes will help us to apply RFM models for the customers of this company, and to calculate the customer lifetime value as well.

3) Data Transformation: No data transformation have been made in the database.

C. RFM and CLV computing

The following tables V and VI show respectively RFM and CLV computing tables.

a) RFM computing: -

CustomerID	Recency	Frequency	Revenue
15905	193	22	114.69
15331	149	44	767.91
17378	23	1	10.95
16153	4	147	3524.27
14301	66	40	1352.04

TABLE V: RFM computing

For each Customer's we get the three scores Recency, Frequency and Revenue.

b) CLV computing: -

CustomerID	CLV
13179	572.31
15884	7.9
16450	1193.81
13015	3240.14

TABLE VI: CLV computing

For each Customer's we get the one CLV score.

D. Clustering

1) Features selection: The inputs for the clustering are : Recency, Frequency, Montarey and CLV.

2) Modelisation: For this study we will use K-mean. K-means clustering is one of the popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes. the main objective of K-means is simple: group similar data together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. K-means algorithm:

- **Input**: k (the number of clusters), D (a data set containing n objects).
- **Output**: A set of k clusters.
- **Method**: Arbitrarily choose k objects from D as the initial cluster centers;
- **repeat**

(a) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(b) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

- **until** no change.

CustomerID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	Segment
15905	193	0	22	0	114.69	0	3	Low-Value
15331	149	1	44	0	767.91	0	4	Low-Value
17378	23	2	1	0	10.95	0	5	Mid-Value
16455	10	2	136	2	2686.37	2	9	High-Value
16153	4	2	147	2	3524.27	2	9	High-Value
14301	66	1	40	0	1352.04	1	5	Mid-Value

TABLE VII: RFM segmentation with RFM scores

3) Cluster Optimisation: The elbow method is a famous method of interpretation and validation of consistency within-cluster analysis designed to help find the appropriate number of clusters in a dataset.

(a) Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from one to ten clusters.
(b) For each k, calculate the total within-cluster sum of square (wss).
(c) Plot the curve of wss according to the number of clusters k.
(d) The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

The elbow Method shows that the optimal cluster number for Recency is k=3.

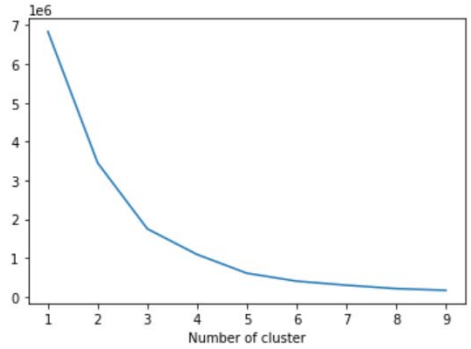


Fig. 3: Elbow result.

4) RFM segmentation and RFM scores: We must assign a score to each customer who has recently transacted with the retail store based on the three important metrics mentioned above: R, F, and M. Then, develop a scoring methodology for segmenting the customer base and applying for various marketing programs. Table VII represent the result RFM segmentation and RFM Scores.

5) CLV Clusters: CLTV assists us in determining how much money to spend on acquiring new customers and retaining existing ones. the following table 4 shows the CLV segments of our data set, the segment 0 is the low, 1 the medium and 2 is the best clv segment. we will use min and max value for every clv cluster to deduce the max ammount we can spend on customer based on company strategy.

LTVCluster	count	mean	std	min	25%	50%	75%	max
0	1856.0	346.594435	190.748366	3.75	184.5975	321.095	495.8200	751.56
1	864.0	1160.668866	281.211715	753.51	915.1100	1125.285	1364.6125	1754.24
2	382.0	2353.514275	408.068752	1758.52	2024.7825	2260.475	2674.6025	3308.73

(a)

Fig. 4: LTV Clusters

E. Analysis

In the analysis, We attempt to put everything together. We combine our feature sets and LTV data and plotting LTV vs RFM overall scores as well as CLV clusters with red, yellow, and green horizontal rectangles. Every customer is represented by a single small circle in three different colors. and based on that we form a few groups with similar characteristics.

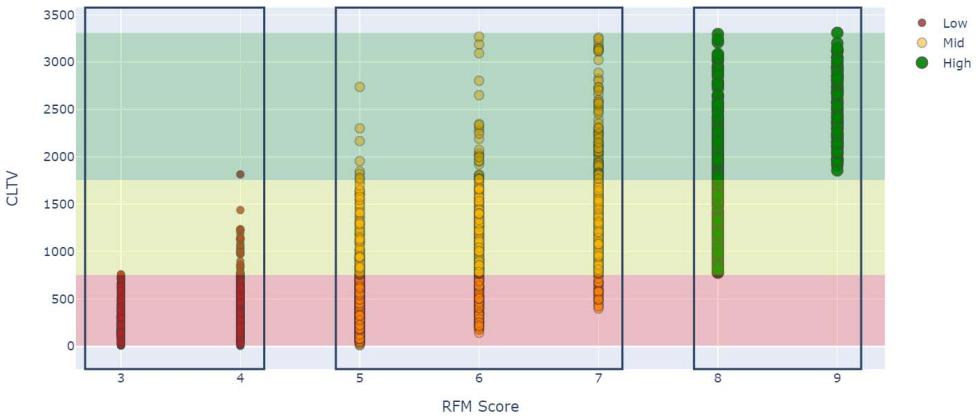
F. Actions

The analysis section provides many groups of customers with similar characteristics, and for each group, we should consider a different strategy to keep them active in our store and purchasing from us more frequently. There are several recommended actions and programs we can implement to achieve those goals, but not every group responds the same way to those actions, so we should define a strategy for each group by analyzing their characteristics and acting accordingly.

V. RESULTS

A. CLV and RFM segments correlation

We attempted to put everything together in Figure 5. The graph below merges our feature sets and LTV data and plots LTV vs RFM overall scores, as well as the CLV clusters with red, yellow, and green horizontal rectangles. Every customer is represented by a single small circle in three different colors: red, yellow, and green, and every customer belong to a specific RFM segment (low, mid and high). The intersection of CLV clusters and RFM Clusters gives us 9 groups of customers with common characteristics, and for each group, we should consider a different strategy to keep them active in our store and buy from us more often. There are several recommended actions and programs we can run to achieve those objectives (advertising, cashback offer, gift coupon ...), but not every group responds to those actions in the same way, so we should define a strategy for each group by analyzing their characteristics and acting accordingly.



(a)

Fig. 5: LTV vs RFM Scores

B. Targeted Actions

It's time to move on to the next step in the process: increasing customer lifetime value. Marketers are always looking for new ways to gain insight into their customers' behavior. What factors influence their purchasing decisions? What drives them? We now have 9 customer segments, each with its own set of characteristics and belonging to one RFM cluster and one CLV cluster. Table VIII represents customer groups with RFM Score Rang and CLV Rang, as well as the characteristics of each group with recommended actions.

VI. Conclusion and Future Directions

In this paper we proposed a customer segmentation model Based on RFM and CLV using unsupervised Clustering. The model analyses customers' transactional data and proposes a set of recommended actions for every segment to enhance the company's profit. Our model has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers.

To encourage repeat purchases, we should offer a loyalty programs and keep in touch with our customers by sending check in and thank you emails. Send coupon codes and other promotional materials help to make customers alive in our store and it's also important to make use of social media to funnel traffic and raise brand awareness.

Future research should carefully consider the potential effects of recommended actions, it will be ideal to apply the model for multiple seasons and measure how much improvement we make in terms of customer loyalty and company revenue.

Segment	RFM Rang	CLV Rang	Characteristics	Recommended Actions
VIP Customers	8-9	1758.52 - 3308.73	Bought recently, buy often and spend the most.	Special prizes, platinum card, assistance, maintain good relationships.
Potential Customers	8-9	753.51 - 1754.24	Recent customers , but spent a good amount and bought more than once.	Start loyalty programs, silver card, recommend other products, assistance, enhance relationships.
New Customers	8-9	3.75 - 751.56	Bought most recently, but not often.	new user promotions, assistance, build relationships.
Loyal Customers	5-7	1758.52 - 3308.73	Spend good money with us often.	Gold Card, recommend promotions, assistance, enhance relationships, Ask for reviews.
Promising Customers	5-7	753.51 - 1754.24	Recent shoppers, but haven't spent much.	provide free trial, assistance, enhance relationships .
Giving up Customers	5-7	3.75 - 751.56	Below average recency, frequency and monetary values. Will lose them if not reactivated.	reconnect with them, recommend promotions and popular products, assistance, rebuild relationships.
Onetime Customers	3-4	1758.52 - 3308.73	Made biggest purchase, but haven't returned for a long time.	Retain them with renewals or newer items; don't let them go to the competition; instead, communicate with them.
Sleeping Customers	3-4	753.51 - 1754.24	Last purchase was long back, low spenders and bought seldomly.	Provide additional related products and exceptional prices, assistance, fix the relationships.
Loosed Customers	3-4	3.75 - 751.56	Lowest recency, frequency and monetary scores (RFM score) and haven't spent much.	Revive interest with a reach out campaign, ignore otherwise.

TABLE VIII: Customers Characteristics and Recommended Actions

REFERENCES

[1] A. Gallo, "the-value-of-keeping-the-right-customers," 2014.

[2] D.-H. Shymko, "https://www.instinctools.com/blog/how-to-avoid-a-bad-customer-experience-in-ecommerce-five-mistakes-to-learn-from/."

[3] C. Butt, H. Hakim, J. Jacobs, and R. Schaffner, "An essential marketing tool in a downturn: Spend management," no. May, 2020.

[4] V. Kumar, R. Venkatesan, T. Bohling, and D. Beckmann, "The power of clv: Managing customer lifetime value at ibm," *Marketing Science*, vol. 27, 2008.

[5] N. Glady, B. Baesens, and C. Croux, "A modified pareto/nbd approach for predicting customer lifetime value," *Expert Systems with Applications*, vol. 36, 2009.

[6] J. W. Scholars and R. Rajagopalan, "Scholarlycommons scholarlycommons a recency-only pareto/nbd a recency-only pareto/nbd," 2018. [Online]. Available: https://repository.upenn.edu/joseph-wharton-scholars/58

[7] H. Hwang, T. Jung, and E. Suh, "An ltv model and customer segmentation based on customer value: A case study on the wireless telecommunication industry," *Expert Systems with Applications*, vol. 26, 2004.

[8] B. A. Gloy, J. T. Akridge, and P. V. Preckel, "Customer lifetime value: An application in the rural petroleum market," *Agribusiness*, vol. 13, 1997.

[9] W. Chaimanowong and T. Ke, "A micro-founded model for clv," *SSRN Electronic Journal*, 2021.

[10] L. Abidar, D. Zaidouni, and A. Ennouaary, "Customer segmentation with machine learning: New strategy for targeted actions," 2020.

[11] B. von Mutius and A. Huchzermeier, "Customized targeting strategies for category coupons to maximize clv and minimize cost," *Journal of Retailing*, vol. 97, pp. 764–779, 12 2021.

[12] C. P. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of k-means algorithm for efficient customer segmentation: A strategy for targeted customer services," 2015. [Online]. Available: www.ijarai.thesai.org

[13] B. von Mutius and A. Huchzermeier, "Customized targeting strategies for category coupons to maximize clv and minimize cost," *Journal of Retailing*, vol. 97, pp. 764–779, 12 2021.

[14] T. V. Binh, N. G. Thy, and H. T. N. Phuong, "Measure of clv toward market segmentation approach in the telecommunication sector (vietnam)," *SAGE Open*, vol. 11, 2021.

[15] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," 2021.

[16] D. Jain and S. S. Singh, "Customer lifetime value research in marketing: A review and future directions," *Journal of Interactive Marketing*, vol. 16, 2002.

[17] kaggle, *onlineretail*, https://www.kaggle.com/datasets/vijayuv/onlineretail.